Optimisation de l'échelle d'observation pour l'annotation d'images

Mathis CORDIER[1], Pejman RASTI[1], Cindy TORRES[2] David ROUSSEAU[1]

[1]LARIS, UMR INRAe-IRHS, Université d'Angers, 62 avenue Notre Dame du Lac, 49000 Angers, France

[2]Vilmorin-Mikado, Rue du Manoir, 49250 La Ménitré, France

david.rousseau@univ-angers.fr

Résumé – L'annotation des données constitue aujourd'hui un goulot d'étranglement au développement de l'apprentissage machine supervisé. En effet, nous disposons de modèles très efficaces et de capacités de calcul suffisamment importantes pour entraîner ces modèles. En revanche, il est compliqué d'obtenir des bases de données annotées suffisamment importantes pour ceux-ci, d'autant plus lorsque l'annotation nécessite d'être faite par un expert du domaine. Dans le cadre de détection de symptômes de maladie sur des plantes, nous avons besoin que ces annotations soient faites par des spécialistes et mettons en place des campagnes d'annotation auprès de ces derniers. Nous cherchons alors un moyen de rendre ces campagnes d'annotation plus efficaces, en réduisant le temps d'annotation des images pour les spécialistes et en diminuant ainsi leur fatigue. Nous montrons dans cette communication l'existence d'une échelle optimale d'observation pour ces annotations. Notre approche empirique est fondée sur l'analyse statistique de signaux issus d'eye-tracker. Elle ouvre la voie à des questionnements plus méthodologiques.

Abstract – Data annotation is currently a bottleneck in the development of Deep Learning. Indeed, we have very efficient models and sufficiently large computing capacities to train these models. On the other hand, it is complicated to obtain sufficiently large annotated databases for them, especially when the annotation needs to be done by a domain expert. In the context of detecting disease symptoms on plants, we need these annotations to be done by specialists and we set up annotation campaigns with them. We are then looking for a way to make these annotation campaigns more efficient, by reducing the annotation time for the specialists and thus reducing their fatigue. We show in this paper the existence of an optimal observation scale for these annotations. Our empirical approach is based on the statistical analysis of eye-tracker signals. It opens the way to more methodological questions.

1 Introduction

La compréhension de la vision humaine est un objet d'étude important dans la communauté signal-image. Objet de fascination pour comprendre la façon dont la nature a sélectionné un système aussi perfectionné pour réaliser des tâches complexes en temps réel, c'est aussi une source d'inspiration pour des traitements numériques. Lorsque les systèmes technologiques évoluent, l'étude de la vision dans ces nouveaux contextes est renouvelée, comme par exemple avec l'arrivée de la télévision en 1956 [1]. À l'heure de l'apprentissage machine automatique, le système visuel humain est sollicité dans ce qui devient une industrie de l'annotation d'images. En effet, une des conséquences de l'usage du calcul intensif pour l'apprentissage machine automatique dans des approches pilotées par la data est le besoin accru d'annotations de qualité des images. Ces annotations sont d'autant plus chronophages à établir que le détail à extraire de l'image est fin ou que le modèle d'apprentissage visé est gourmand en nombre de paramètres à ajuster (deep learning en particulier). Ainsi, établir la vérité terrain pour une segmentation prend en général plus de temps que pour une détection d'objets ou une classification. Cette annotation est également plus longue lorsque les images sont de grande dimension (coupes histologiques, vues de télédétection) ou que les items d'intérêt sont relativement petits dans l'image, c'est à dire que l'on cherche *une aiguille dans une botte de foin*. Dans cette situation, une approche peut consister à découper les grandes images en images plus petites afin de réduire la complexité de la tâche dans chaque imagette. La question d'un compromis apparaît. Jusqu'où faut-il découper les images ? En effet, un découpage très fin permettra de faire apparaître les objets cherchés ou l'absence de ces objets plus aisément. Mais, bien sûr, le temps d'annotation sera d'autant plus long qu'il y a davantage d'images à observer. Un optimum doit exister et il est d'un intérêt pratique réel. Nous proposons de montrer cette existence sur un cas d'étude en imagerie des plantes et discutons les moyens possibles pour aller vers une détermination automatique a priori de cet optimum.

2 Cas d'étude

Pour illustration, nous considérons une tâche informationnelle de détection de symptômes de maladies sur plantes à un stade précoce dans des images couleur classiques. Les images acquises sur un piquet connecté sont prises à distances égales de la scène. Leur résolution est de 4600x3264 pixels et les symptômes constituent des petites taches marrons d'environ 50x50 (voir Fig.1). On appelle échelle d'observation la quantité par laquelle on découpe horizontalement et verticalement notre image originale pour former des imagettes à annoter. Comme

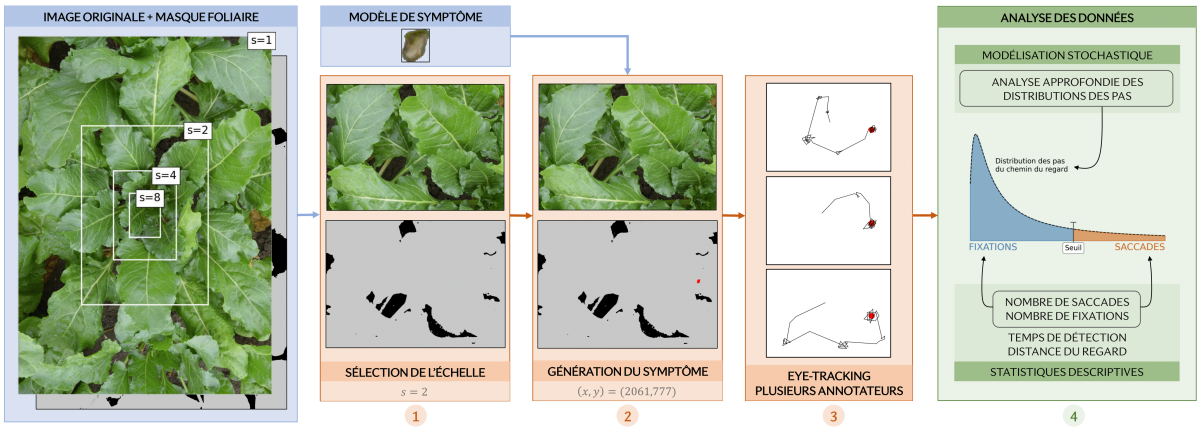

FIGURE 1 – Pipeline de génération, acquisition et traitement de nos données. Au départ du pipeline, nous disposons de l'image originale, son masque foliaire et un modèle de symptôme. Pour générer une image, on sélectionne aléatoirement une échelle parmi Ω_s et une des imagettes dans le quadrillage associé en ①. L'imagette sélectionnée peut ensuite subir une rotation. En ②, on ajoute notre modèle de symptôme sur une position aléatoirement choisie dans le masque foliaire. En ③, on procède à l'annotation de l'image générée pour tous les annotateurs. Enfin, on analyse les marches aléatoires 2D obtenues suite à l'eye-tracking en ④.

représenté en Fig.1, une échelle s grande équivaut donc à un zoom important dans la scène. Cela se traduit par une réduction de la complexité des imagettes et une augmentation de la quantité d'imagettes à annoter. Pour notre étude, nous utilisons l'ensemble d'échelles d'observation $\Omega_s = \{1, 1.5, 2, 3, 4, 6, 8\}$.

2.1 Jeu de données

Afin de disposer de vérités terrain pré-établies, nous générons des images de synthèse en plaçant aléatoirement un symptôme dans des images originales qui en sont dépourvues. Tout le processus de construction de notre jeu de données est illustré en Fig.1. Le générateur permet de simuler des symptômes seulement sur les zones de l'image correspondant aux feuilles de la plante. Au total, 420 images sont générées dans la base, soit 60 par échelle avec l'ensemble des échelles étudiées Ω_s.

2.2 Eye-tracking

Pour annoter les images, nous utilisons un système d'eye-tracking comme précédemment utilisé en [2]. Ce capteur enregistre sans contact la position du regard lors de l'inspection de l'image par l'annotateur. L'expérience a été réalisée sur un écran d'une résolution de 1920×1080 pixels à l'aide d'un eye-tracker binoculaire à distance de type SMI cadencé à 120 Hz. Cet outil nous permet de capter avec précision la position du regard sur l'écran lors de la recherche du symptôme. Nous avons répété l'expérience avec 3 annotateurs pour un total de 658 annotations. Dans le protocole d'acquisition, les annotateurs prennent connaissance du modèle de symptôme à détecter en amont de la phase d'annotation. Afin de valider la détection du symptôme lors de cette phase, il est demandé aux annotateurs de le fixer pendant 3 secondes au moment de sa détection.

3 Analyse des données

Les mouvements de l'oeil humain sont classiquement décrits sous forme d'alternance de saccades et de fixations [3, 4]. Les saccades sont des sauts très rapides effectués par le regard dans la scène et les fixations correspondent à des temps plus longs sur lequel le regard se stabiliser sur un endroit de la scène. Dans une première partie, nous analysons les temps de détection, la distance effectuée par le regard et le nombre de saccades et fixations pour chaque échelle. La classification entre les saccades et les fixations, représentée en Fig.1④ est réalisée directement par BeGaze, le logiciel d'eye-tracking utilisé. Dans une seconde partie, nous nous affranchirons de cette binarisation pour analyser directement les distributions des pas de la marche aléatoire produite par le déplacement du regard sur l'écran.

3.1 Saccades et fixations

Nous décrivons pour commencer les trajectoires du regards des annotateurs pour chaque échelle s au moyen de 4 variables : le temps de détection du symptôme Δt_s en secondes, la distance parcourue d_s en pixels, le nombre de fixations $\mathcal{F}_s$ et le nombre de saccades $\mathcal{S}_s$ effectués par le regard lors de la recherche. Afin d'analyser comment les distributions diffèrent selon les facteurs échelle et annotateur, nous appliquons une ANOVA sur nos données.

	p-value			
	Δt	d	$\mathcal{F}$	$\mathcal{S}$
Échelle	$< 10^{-16}$	$< 10^{-16}$	$< 10^{-16}$	$< 10^{-16}$
Annotateur	0.249	0.611	0.201	0.608
Échelle + Annotateur	0.987	0.495	0.989	0.776

TABLE 1 – ANOVA à 2 facteurs : l'échelle s et l'annotateur.

Les résultats sont présentés en Tab.1. Ils nous permettent de rejeter clairement l'hypothèse nulle d'égalité des moyennes des distributions selon les échelles. De plus, ils confirment l'hypothèse selon laquelle les annotateurs se comportent de manière sensiblement identiques durant l'expérience. Ainsi, nous étudierons désormais les signaux obtenus tout annotateur confondu. Les variables d'intérêt présentées précédemment ne tiennent pas compte du fait que nous avons s^2 imagettes à l'échelle s. En effet, cela correspond à un découpage de l'image originale selon un quadrillage sxs. Nous introduisons alors les variables d'intérêt normalisées définies par $\Delta t_s^* = \Delta t_s \cdot s^2$, $d_s^* = d_s \cdot s^2$, $\mathcal{F}_s^* = \mathcal{F}_s \cdot s^2$ et $\mathcal{S}_s^* = \mathcal{S}_s \cdot s^2$. Chacune de ces variables normalisées représente respectivement le temps total nécessaire pour détecter le symptôme pour toutes les imagettes à l'échelle s, la distance totale effectuée par le regard à l'échelle s, le nombre de fixations total et le nombre de saccades total.

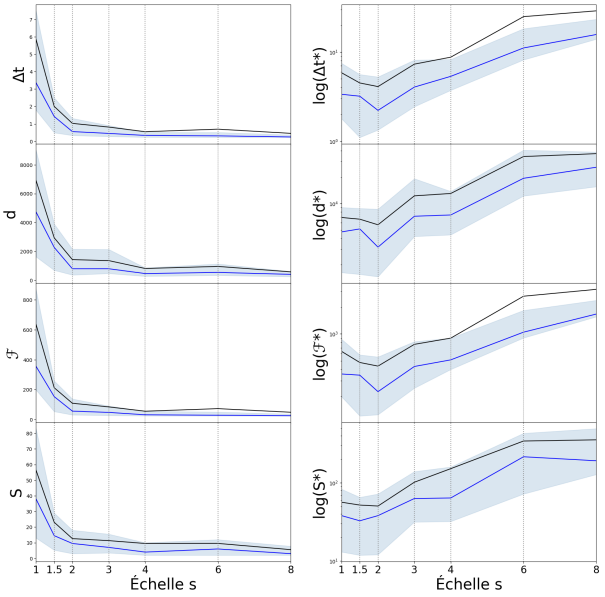

FIGURE 2 – Les courbes noires représentent l'évolution de la moyenne et les bleues celle de la médiane. Les surfaces bleues délimitent l'intervalle entre le premier et le troisième quartile. À gauche, représentation de l'évolution de ces indicateurs sur nos variables d'intérêt non normalisées en fonction de l'échelle d'observation. À droite, représentation de ces indicateurs sur nos variables d'intérêt normalisées (échelle logarithmique) en fonction de l'échelle d'observation. On remarque que la moyenne est toujours supérieure à la médiane. Cela est due à l'asymétrie des distributions qui sont décalées à gauche de la médiane et possèdent un coefficient d'asymétrie (skewness) positif comme visible sur la Fig.1④.

Comme illustré sur la Fig.2, les variables normalisées suivent une évolution non monotone qui passe par un optimum pour l'échelle $s = 2$. Nous montrons ainsi expérimentalement l'existence d'une échelle optimale d'observation pour l'annotation des images. En observant les images de notre base de données à l'échelle optimale $s = 2$, on obtient ici le meilleur compromis avec 4 fois plus d'images à annoter mais une forte réduction de la complexité des images. Comme représenté en Tab.2, nous avons un gain très important de 30% du temps global d'annotation. Il est à noter que nous n'incluons pas ici le temps nécessaire pour découper les images car il est négligeable sur nos images par rapport au temps nécessaire pour l'inspection et l'enregistrement de l'annotation. Ce point serait sans doute à considérer pour des images de très grande dimension.

	Δt^*	d^*	$\mathcal{F}^*$	$\mathcal{S}^*$
μ	-30%	-17.5%	-31.5%	-10.6%
σ	-31.1%	-24.2%	-29.9%	-25.8%

TABLE 2 – Tableau indicatif du gain sur nos variables d'intérêts normalisées pour un passage à l'échelle optimale $s = 2$. En première ligne, le gain sur la moyenne μ, en deuxième ligne le gain sur l'écart type σ.

Une seconde observation sur les variables d'intérêt non normalisées montre l'apparition d'un plateau à partir de l'échelle 4. Cela signifie que lorsque l'on zoome suffisamment, le temps d'analyse semble ne plus évoluer (i.e. ne plus diminuer ici). Cette observation est cohérente avec l'existence de la vision pré-attentive [5] dans laquelle la détection d'objets aberrants (ici le symptôme) devient inconsciente si ce qui distingue ces objets du fond est suffisamment homogène spatialement. Pour approfondir cette observation, nous proposons dans la section suivante une modélisation stochastique des marches aléatoires enregistrées du regard aux différentes échelles d'observation.

3.2 Modélisation stochastique

Dans cette partie, nous ne réalisons pas de seuillage entre saccade et fixation. Nous analysons directement les processus de la longueur des pas issus des trajectoires enregistrées à chaque échelle d'observation. Pour ce faire, sur la Fig.3A, nous traçons les fonctions de répartition des distributions pour chaque échelle d'observation. Nous y montrons également les résultats des tests de Kolmogorov-Smirnof qui viennent valider l'interprétation visuelle sur les différences entre les fonctions de répartition. Ces résultats traduisent une modification du comportement du regard en fonction de l'échelle d'observation. Plus précisément, l'évolution de la fonction de répartition montre qu'une augmentation de l'échelle correspond à une augmentation des petits pas par rapport aux grands. Cela corrobore l'idée qu'une annotation devienne plus simple lorsque l'échelle devient plus grande. Enfin, ces évolutions stagnent pour des plus grandes échelles (courbes pour l'échelle 6 et 8 très proches l'une de l'autre et KS test non significatif). Là aussi, cette observation est cohérente avec la stagnation enregistrée dans l'approche avec seuillage entre saccade et fixation. Sur la Fig.3B, les fonctions d'autocorrélation des marches aléatoires sont fournies. À décalage faible, on y distingue des corrélations plus importantes sur les grandes échelles. Lorsqu'on augmente ce décalage, les corrélations deviennent plus importantes sur les

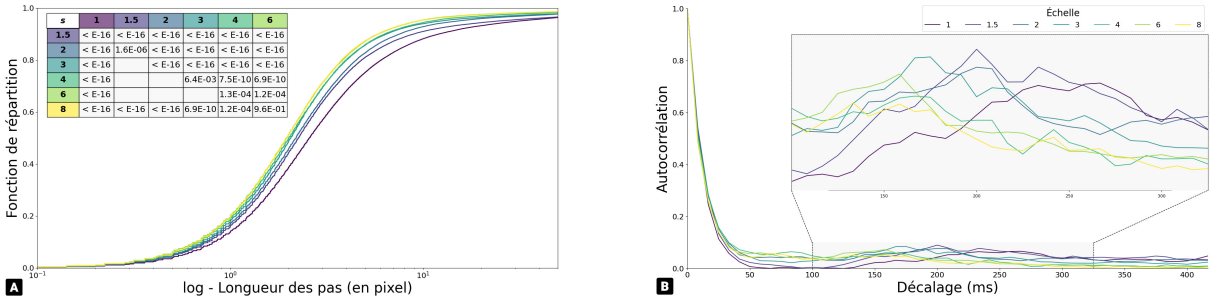

FIGURE 3 – Étude de la longueur des pas des trajectoires du regard pour chaque échelle s. À gauche, les fonctions de répartition des distributions englobant l'ensemble des longueurs des pas pour chaque échelle. Le tableau représente les p-values des KS tests entre les distributions de chaque échelle. On rejette l'hypothèse nulle d'égalité des lois pour toutes les comparaisons d'échelle à échelle hors 6-8. À droite, la fonction d'autocorrélation moyennant l'ensemble des processus pour chaque échelle.

petites échelles. Cette inversion, observée autour de 170 ms, révèle que les cycles formés par l'alternance des saccades et des fixations sont d'autant plus courts que l'échelle est élevée. Cela peut être expliqué par des images moins complexes quand l'échelle d'observation est grande.

4 Conclusion et perspectives

L'expérience décrite dans cette communication montre l'existence d'une échelle optimale d'observation pour l'annotation d'images où le temps nécessaire pour l'annotation est minimal. Aux petites échelles, peu d'images sont à annoter mais le temps d'inspection est très long. Aux grandes échelles, la tâche se simplifie et on peut même atteindre des régimes semblables à la vision pré-attentive tels que décrits dans la théorie de Gestalt [5]. Le gain de temps par rapport au choix par défaut d'annotation d'une image entière dans notre cas d'étude est substantiel puisqu'il s'élève à environ 30%. Si l'illustration était donnée en imagerie pour les plantes, les problématiques posées sont d'intérêt général en traitement du signal-image. Ces résultats préliminaires appellent à des extensions.

Sur le plan expérimental, il serait intéressant de reproduire ces observations dans un cadre psycho-visuel calibré avec un plus grand nombre d'annotateurs et en faisant varier le nombre d'objets à détecter ainsi que leur similarité avec le fond. Il serait également utile d'apporter une modélisation statistique à la manière de [5] sur ce problème spécifique d'annotation en analysant finement la façon dont les trajectoires de l'oeil évoluent par rapport à des modèles de marches aléatoires à queues lourdes tels les vols de Lévy par exemple [6, 7].

Sur le plan méthodologique, il serait intéressant d'étudier si une échelle optimale d'observation pourrait être définie théoriquement à la manière de précédentes études dans des images bruitées minimalistes [8]. On peut par exemple penser aux métriques de clutter [9] qui peuvent servir à mesurer objectivement la similarité entre un objet et le fond qui l'entoure. Une estimation des échelles spatiales de cartes de clutter devrait permettre de définir des échelles auxquelles la vision pré-attentive peut se mettre en action réduisant ainsi le temps d'annotation [10].

5 Remerciements

M. Cordier remercie l'ITB et la contribution financière du CASDAR du Ministère de l'Agriculture et de l'Alimentation.

Références

[1] G. Sziklai. Some studies in the speed of visual perception. *IEEE Transactions on Information Theory*, 1956.

[2] S. Samiei, P. Rasti, P. Richard, G. Galopin, and D. Rousseau. Toward Joint Acquisition-Annotation of Images with Egocentric Devices for a Lower-Cost Machine Learning Application to Apple Detection. *Sensors*, 2020.

[3] R.J. Krauzlis. The control of voluntary eye movements : new perspectives. *The Neuroscientist : A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, 2005.

[4] J.M. Henderson. Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 2003.

[5] A. Desolneux, L. Moisan, and J-M. Morel. *From Gestalt theory to image analysis : a probabilistic approach*. Springer Science & Business Media, 2007.

[6] P. Barthelemy, J. Bertolotti, and D. Wiersma. A Lévy flight for light. *Nature*, 2008.

[7] G.M. Viswanathan, F. Bartumeus, S.V. Buldyrev, J. Catalan, U.L. Fulco, S. Havlin, M.G.E da Luz, M.L. Lyra, E.P. Raposo, and H. Eugene Stanley. Lévy flight random searches in biological phenomena. *Physica A : Statistical Mechanics and its Applications*, 2002.

[8] A. Delahaies, D. Rousseau, and F. Chapeau-Blondeau. Joint acquisition-processing approach to optimize observation scales in noisy imaging. *Optics letters*, 2011.

[9] S.R. Rotman, G. Tidhar, and M.L. Kowalczyk. Clutter metrics for target detection systems. *IEEE Transactions on Aerospace and Electronic Systems*, 1994.

[10] Y. Semizer and M. Michel. Natural image clutter degrades overt search performance independently of set size. *Journal of vision*, 2019.